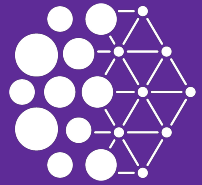
Question Answering in Realistic Visual Environments: Challenges and Approaches

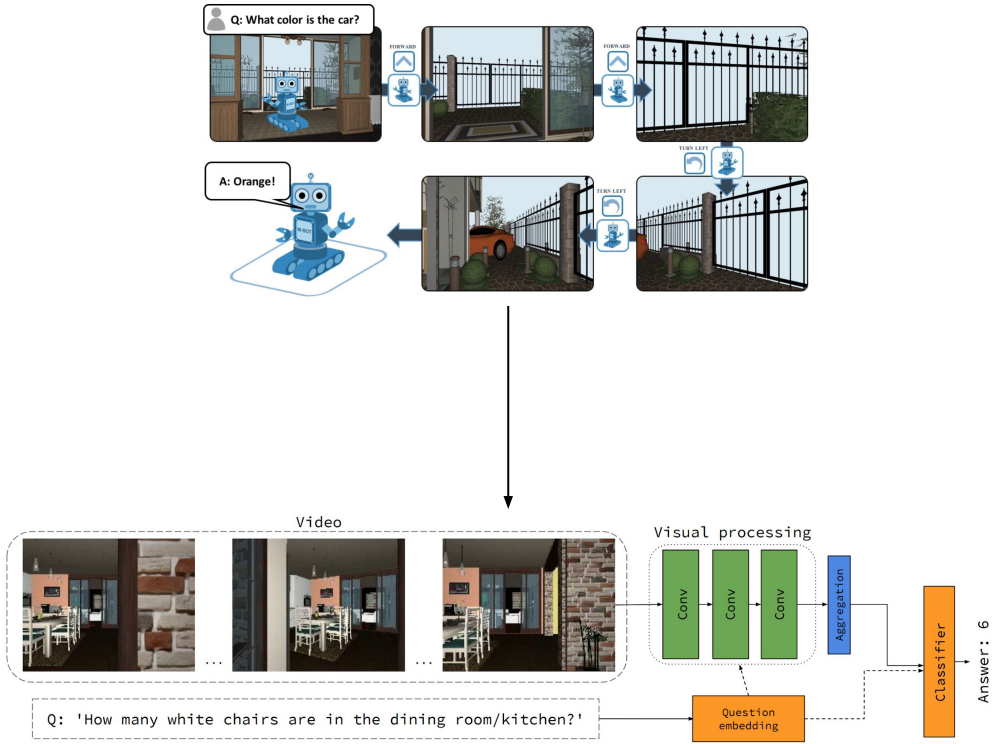
Cătălina Cangea - Mila Tea Talk, Oct 25th 2019

UNIVERSITY OF CAMBRIDGE



Talk outline

1. VQA and embodied agents
2. EQA & related tasks
3. Existing work
4. Perspective shift
    - VideoNavQA task and benchmark
    - Generalized VQA models
    - Initial results

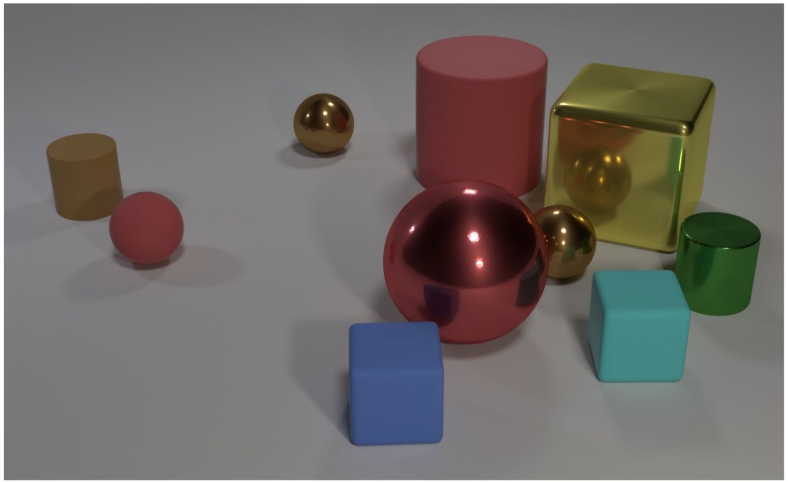


Visual reasoning: the problem space

Visual reasoning

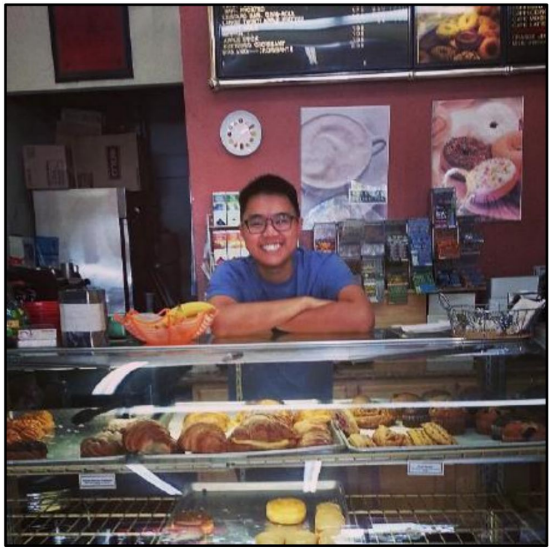
- Vast literature on VQA
- Many benchmarks
- Some task examples
  - CLEVR: small, synthetic concept distribution, functional program-based questions
  - VQA: real-world distribution, natural language questions
  - GQA: real-world distribution, functional program-based questions

CLEVR (Johnson et al., 2016)



Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
Q: **How many** objects are **either small cylinders** or **red** things?

VQA (Antol et al., 2015)



What kind of store is this?

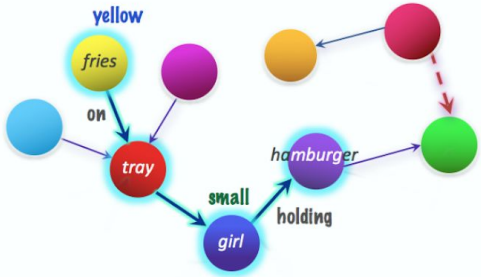
Is the display case as full as it could be?

GQA (Hudson & Manning, 2018)

Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?
Program: **Select:** <dobject> ⟶ **Choose** <type>: <attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger ⟶ **Relate:** girl,holding ⟶ **Filter** size: small ⟶ **Relate:** object, left ⟶ **Filter** color: red ⟶ **Relate:** food,on ⟶ **Choose** color: yellow | brown



Graph Normalization

- Ontology construction
- Edge Pruning
- Object Augmentation
- Global Properties

Question Generation

- Patterns Collection
- Compositional References
- Decoys Selection
- Probabilistic Generation

Sampling and Balancing

- Distribution Balancing
- Type-Based Sampling
- Deduplication

Entailments Relations

- Functional Programs
- Entailment Relations
- Recursive Reachability

New Metrics

- Consistency
- Validity & Plausibility
- Distribution
- Grounding

VQA < ...

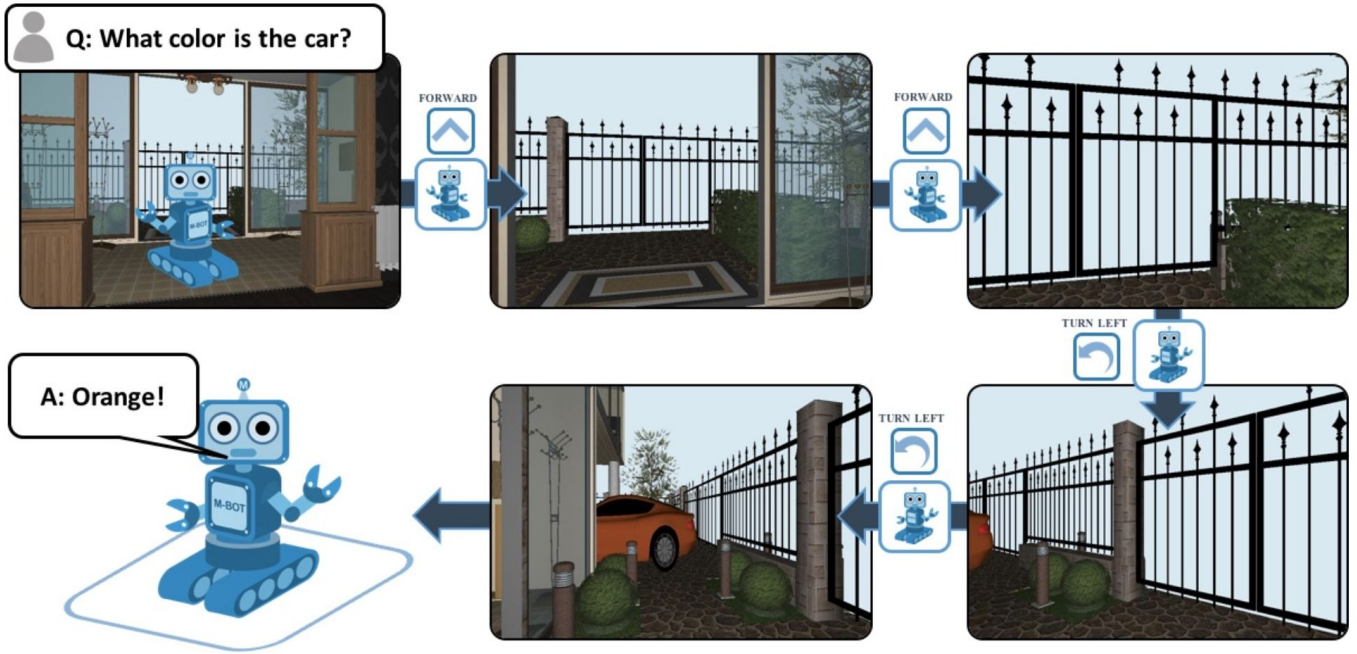
- We can do **reasonably** well on existing tasks
  - Some benchmarks “solved” (99.9...% on CLEVR)
  - Issues still to be addressed
- Still, VQA tasks involve a single image!
- What about more complex settings?
  - Require interaction via embodied agents
  - “*In artificial intelligence, an **embodied agent** [...] is an **intelligent** agent that **interacts with the environment through a physical body** within that environment.*” (Wikipedia)

Embodied agents in the (relative) wild

- Several recently proposed tasks
  - EQA
  - IQA
  - Habitat
  - Vision-and-Language Navigation
- Require an agent to **act and reason** in a **rich, realistic, 3D environment** in order to **answer the given question**

Embodied agent tasks

Embodied QA (Das et al., 2018)

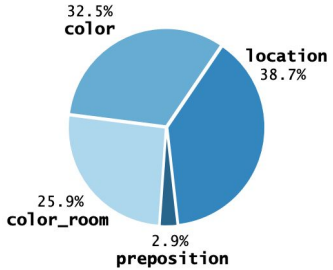


Embodied QA

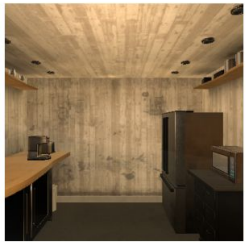
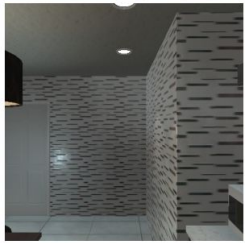
EQA v1

location:	*'What room is the <OBJ> located in?'*
color:	*'What color is the <OBJ>?'*
color_room:	*'What color is the <OBJ> in the <ROOM>?'*
preposition:	*'What is <on/above/below/next-to> the <OBJ> in the <ROOM>?'*
existence:	*'Is there a <OBJ> in the <ROOM>?'*
logical:	*'Is there a(n) <OBJ1> and a(n) <OBJ2> in the <ROOM>?'*
count:	*'How many <OBJS> in the <ROOM>?'*
room_count:	*'How many <ROOMS> in the house?'*
distance:	*'Is the <OBJ1> closer to the <OBJ2> than to the <OBJ3> in the <ROOM>?'*

	Environments	Unique Questions	Total Questions
train	643	147	4246
val	67	104	506
test	57	105	529



Interactive QA (Gordon et al., 2017)

Question and answer	Initial Image	Scene View
Q: Is there bread in the room? A: No		
Q: How many mugs are in the room? A: 3		
Q: Is there a tomato in the fridge? A: Yes		

Habitat Challenge (Savva et al., 2019)

Room2Room/V&LN (Anderson et al., 2017)

Exit the bathroom. Turn left and exit the room using the door on the left. Wait there.

Challenges

What is needed for complex-setting QA?

- **“act”**
  - Learn to explore and navigate
  - Sometimes even interact with the environment
- **“reason...rich, realistic, 3D environment”**
  - Visual processing of the setting
  - Rich stream of visual data arising from the navigation aspect
- **“answer the given question”**
  - **Interpret** the natural language-like conditioning signal
  - **Filter** the visual concepts, select relevant information
  - **Relate** concepts

What is needed for complex-setting QA?

- **“act”**
  - Learn to explore and navigate
  - Sometimes even interact with the environment

RL, planning, IL

- **“reason...rich, realistic, 3D environment”**
  - Visual processing of the setting
  - Rich stream of visual data arising from the navigation aspect

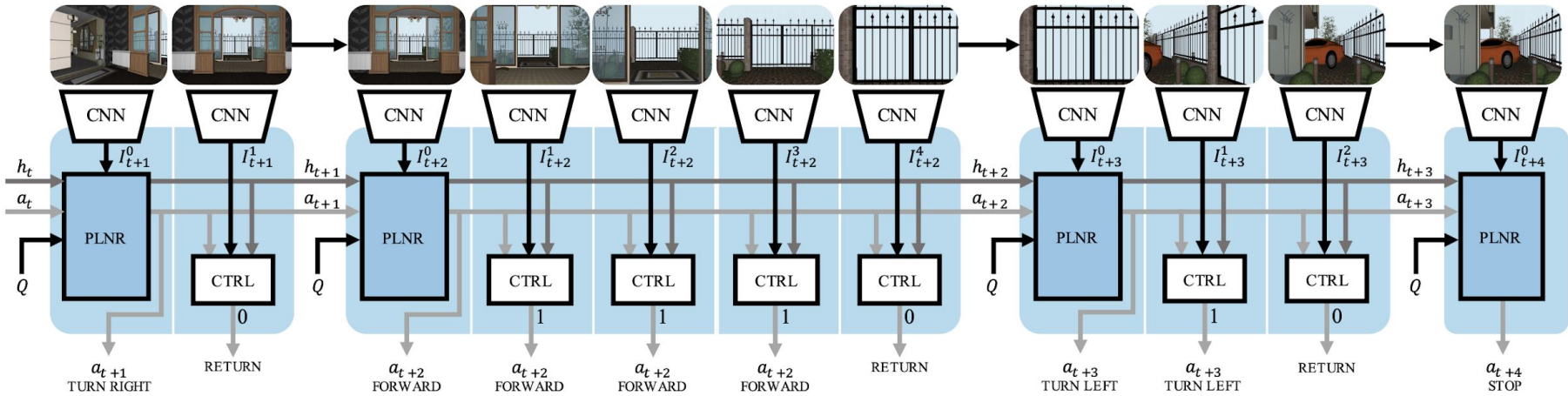
Segmentation, object detection, relation inference

- **“answer the given question”**
  - **Interpret** the natural language-like conditioning signal
  - **Filter** the visual concepts, select relevant information
  - **Relate** concepts

Multimodal processing (language + visual features), powerful conditioning of visual features

Some existing work

EQA initial approach (Das et al., 2018)



EQA initial results

		Navigation																		QA		
		$\mathbf{d_T}$			$\mathbf{d_\Delta}$			$\mathbf{d_{min}}$			$\%\mathbf{r_T}$			$\%\mathbf{r_{\hookleftarrow}}$			$\%\mathbf{stop}$			MR		
		T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}
Baselines	Reactive	2.09	2.72	3.14	-1.44	-1.09	-0.31	0.29	1.01	1.82	50%	49%	**47%**	52%	53%	48%	-	-	-	3.18	3.56	3.31
	LSTM	1.75	2.37	2.90	-1.10	-0.74	-0.07	0.34	1.06	2.05	55%	53%	44%	59%	57%	50%	80%	75%	80%	3.35	3.07	3.55
	Reactive+Q	1.58	2.27	2.89	-0.94	-0.63	-0.06	0.31	1.09	1.96	52%	51%	45%	55%	57%	**54%**	-	-	-	3.17	3.54	3.37
	LSTM+Q	1.13	2.23	2.89	-0.48	-0.59	-0.06	0.28	0.97	1.91	**63%**	53%	45%	64%	59%	**54%**	80%	71%	68%	3.11	3.39	3.31
Us	PACMAN+Q	**0.46**	**1.50**	**2.74**	**0.16**	**0.15**	**0.12**	0.42	1.42	2.63	58%	54%	45%	60%	56%	46%	**100%**	**100%**	**100%**	**3.09**	3.13	3.25
	PACMAN-RL+Q	1.67	2.19	2.86	-1.05	-0.52	0.01	**0.24**	**0.93**	**1.94**	57%	**56%**	45%	**65%**	**62%**	52%	32%	32%	24%	3.13	**2.99**	**3.22**
Oracle	HumanNav*	0.81	0.81	0.81	0.44	1.62	2.85	0.33	0.33	0.33	86%	86%	86%	87%	89%	89%	-	-	-	-	-	-
	ShortestPath+VQA	-	-	-	0.85	2.78	4.86	-	-	-	-	-	-	-	-	-	-	-	-	3.21	3.21	3.21

Neural Modular Control (Das et al., 2018)

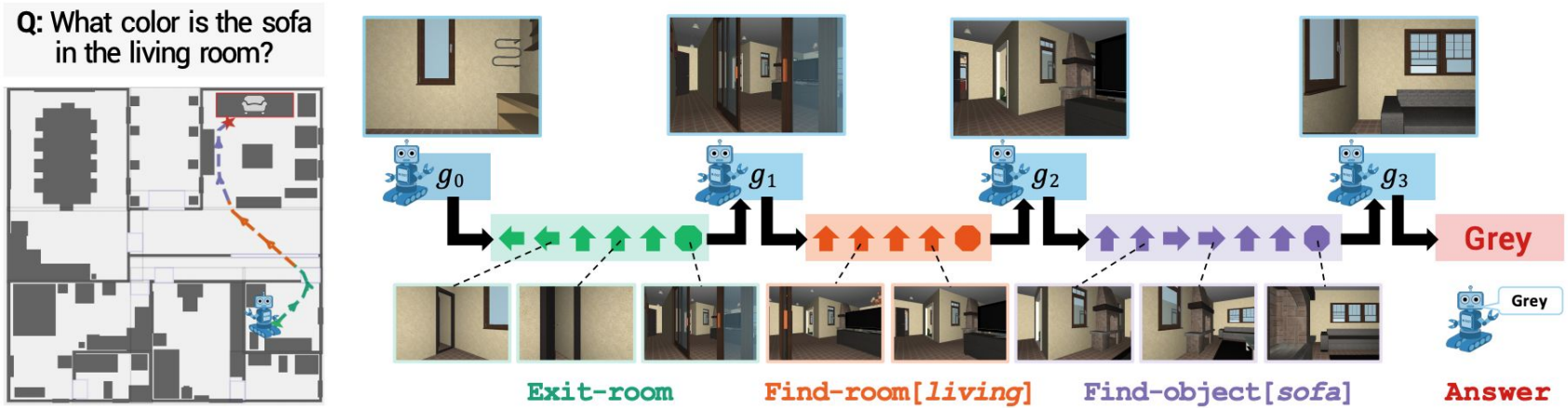


Figure 1: We introduce a hierarchical policy for Embodied Question Answering. Given a question ("What color is the sofa in the living room?") and observation, our master policy predicts a sequence of subgoals – `Exit-room`, `Find-room[living]`, `Find-object[sofa]`, `Answer` – that are then executed by specialized sub-policies to navigate to the target object and answer the question ("Grey").

Neural Modular Control

- IL (**expert** trajectories)
- RL (A3C)
- 2-level policy hierarchy (GRUs)
- 4 **hand-crafted** low-level policies

Subgoals ⟨Tasks, Arguments⟩. As mentioned above, each subgoal is factorized into a task and an argument $g = \langle g_{\text{task}}, g_{\text{argument}} \rangle$. There are 4 possible tasks – `exit-room`, `find-room`, `find-object`, and `answer`. Tasks `find-object` and `find-room` accept as arguments one of the 50 objects and 12 room types in EQA v1 dataset [1] respectively; `exit-room` and `answer` do not accept any arguments. This gives us a total of $50 + 12 + 1 + 1 = 64$ subgoals.

⟨exit-room, none⟩, ⟨answer, none⟩,	} 0 args
⟨find-object, couch⟩, ⟨find-object, cup⟩, ..., ⟨find-object, xbox⟩,	} 50 args
⟨find-room, living⟩, ⟨find-room, bedroom⟩, ..., ⟨find-room, patio⟩.	} 12 args

Neural Modular Control results on EQAv1

	Navigation									QA		
	$\mathbf{d_0}$ (For reference)			$\mathbf{d_T}$ (Lower is better)			$\mathbf{d_\Delta}$ (Higher is better)			**accuracy** (Higher is better)		
	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}
PACMAN (BC) [1]	1.15	4.87	9.64	1.19	4.25	8.12	-0.04	0.62	1.52	48.48%	40.59%	39.87%
PACMAN (BC+REINFORCE) [1]	1.15	4.87	9.64	**1.05**	4.22	8.13	**0.10**	0.65	1.51	50.21%	42.26%	40.76%
NMC (BC)	1.15	4.87	9.64	1.44	4.14	8.43	-0.29	0.73	1.21	43.14%	41.96%	38.74%
NMC (BC+A3C)	1.15	4.87	9.64	1.06	**3.72**	**7.94**	0.09	**1.15**	**1.70**	**53.58%**	**46.21%**	**44.32%**

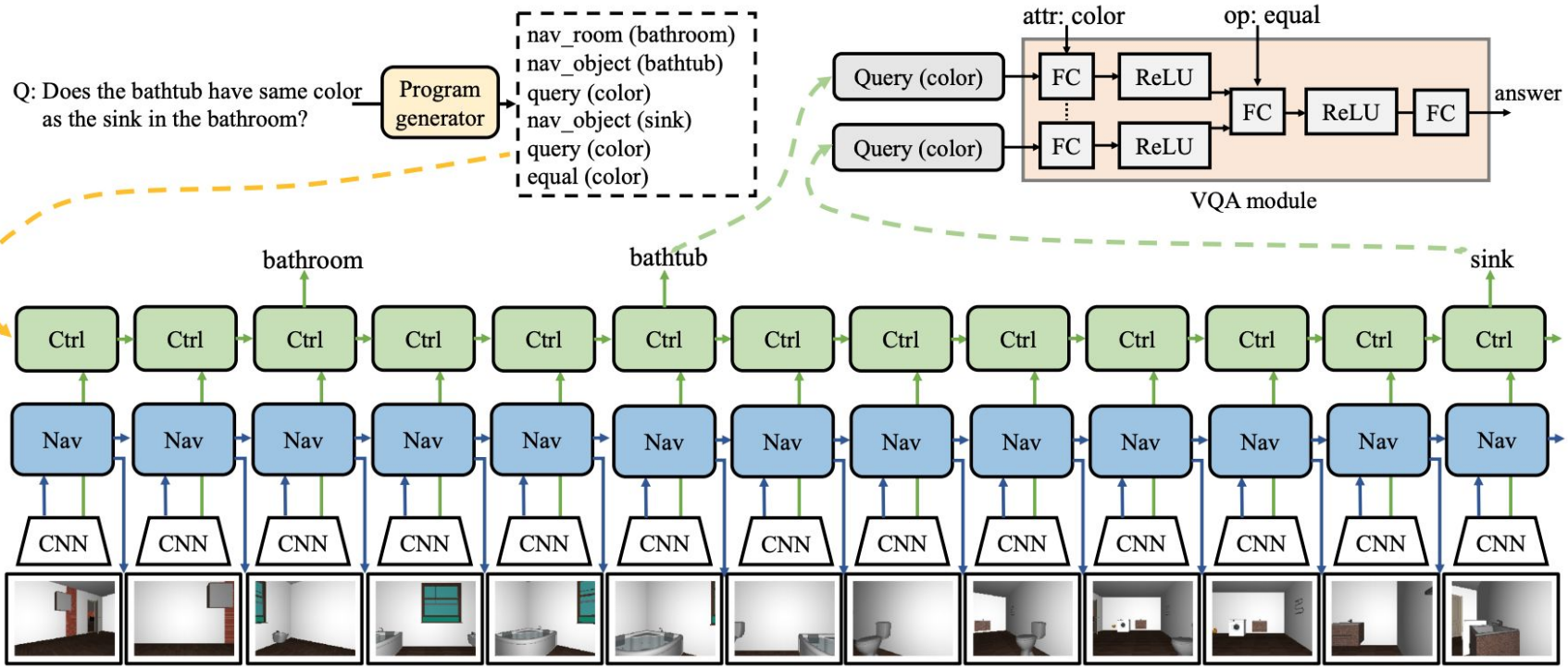
“Blindfold” baselines (Anand et al., 2018)

	T_{10}	T_{20}	T_{50}	T_{any}
Navigation + VQA				
PACMAN (BC) [5]	48.48	40.59	39.87	N/A
PACMAN (BC+REINFORCE)[5]	50.21	42.26	40.76	N/A
NMC (BC) [6]	43.14	41.96	38.74	N/A
NMC (BC+A3C) [6]	**53.58**	46.21	44.32	N/A
Question only				
Majority	17.15	17.15	17.15	17.15
Nearest Neighbor Answer	48.45	48.45	48.45	48.45
BOW	50.34	**50.34**	**50.34**	**50.34**
PACMAN Q-only (LSTM) (*)	46.07	46.07	46.07	46.07

Multi-Target EQA (Yu et al., 2019)

	Question Type	Template
EQA-v1	location	*"What room is the <OBJ> located in?"*
	color	*"What color is the <OBJ>?"*
	color_room	*"What color is the <OBJ> in the <ROOM>?"*
	preposition	*"What is <on/above/below/next-to> the <OBJ> in the <ROOM>?"*
MT-EQA	object_color_compare_inroom	*"Does <OBJ1> share same color as <OBJ2> in <ROOM>?"*
	object_color_compare_xroom	*"Does <OBJ1> in <ROOM1> share same color as <OBJ2> in <ROOM2>?"*
	object_size_compare_inroom	*"Is <OBJ1> bigger/smaller than <OBJ2> in <ROOM>?"*
	object_size_compare_xroom	*"Is <OBJ1> in <ROOM1> bigger/smaller than <OBJ2> in <ROOM2>?"*
	object_dist_compare	*"Is <OBJ1> closer than/farther from <OBJ2> than <OBJ3> in <ROOM>?"*
	room_size_compare	*"Is <ROOM1> bigger/smaller than <ROOM2> in the house?"*

Multi-Target EQA



Multi-Target EQA

		Object Navigation					Room Navigation		EQA				
		d_T	d_Δ	h_T	IOU_T^r	$\%stop_o$	$\%r_T$	$\%stop_r$	ep_len	$\%_{easy}$	$\%_{medium}$	$\%_{hard}$	$\%_{overall}$
1	Nav+cVQA	5.41	-0.64	0.19	0.15	36	34	60	153.13	58.42	53.29	51.46	53.24
2	Nav(RL)+cVQA	3.80	0.10	0.33	**0.30**	46	40	62	144.80	67.57	55.91	53.28	57.40
3	Nav+Ctrl+cVQA	5.25	-0.56	0.20	0.18	36	37	70	145.20	59.73	53.48	49.04	54.44
4	Nav(RL)+Ctrl+cVQA	**3.60**	**0.16**	**0.33**	0.29	**48**	**43**	**72**	**127.71**	**72.22**	**59.97**	**54.92**	**61.45**

Escape the bottleneck?

Complexity!

- **“act”**
  - Learn to explore and navigate
  - Sometimes even interact with the environment
- **“reason...rich, realistic, 3D environment”**
  - Visual processing of the setting
  - Rich stream of visual data arising from the navigation agent
- **“answer the given question”**
  - **Interpret** the natural language-like conditioning signal
  - **Filter** the visual concepts, select relevant information
  - **Relate** concepts

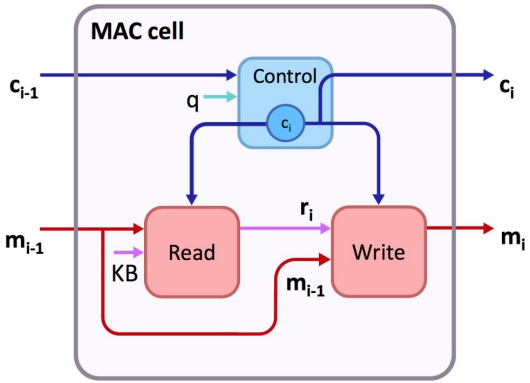
RL, planning, IL

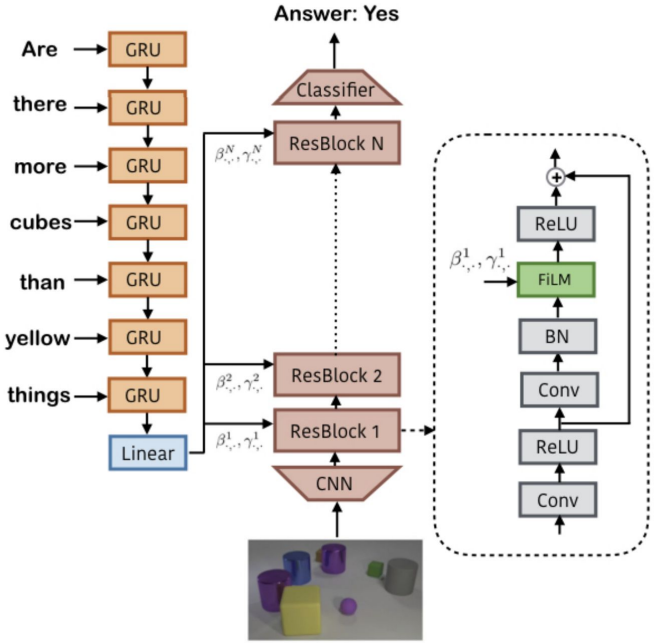
Segmentation, object detection, relation inference

Multimodal processing (language + visual features), powerful conditioning of visual features

A different perspective

- Integrating RL, vision, language and reasoning is hard!
- Can we tackle EQA from a more favorable angle?
  - Perhaps via established visual reasoning paradigms...





Embodied QA: alternative take

- **Remove** navigation aspect completely
  - RL, IL, planning not required anymore
- **Increase the difficulty of reasoning tasks**
  - Shift the focus to interpreting the visual stream
- Provide an **almost-ideal trajectory** of the agent in the environment
  - All the information necessary to answer the question can be found on the path
  - But it might not be the shortest path

Embodied QA: alternative take

- **EQA - navigation + advanced reasoning**
- Input:
  - **more complex question** about the environment
  - **video of agent's trajectory**
- Output: answer

A new task: VideoNavQA

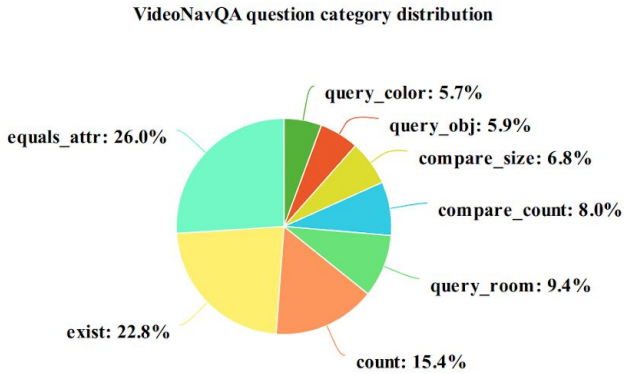
VideoNavQA: Bridging the Gap between Visual and Embodied Question Answering
Cătălina Cangea, Eugene Belilovsky, Pietro Liò, Aaron Courville (BMVC 2019, ViGIL 2019)

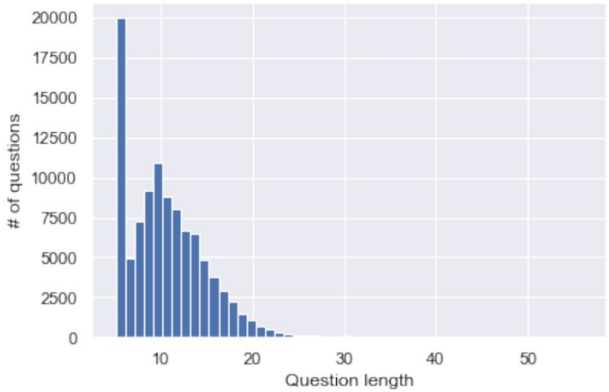
Task overview

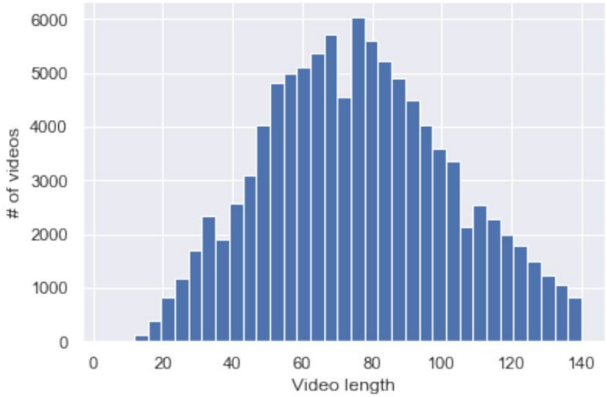
VideoNavQA dataset

- Aim to help study the feasibility of EQA-style tasks
- Generated using House3D
- 100k samples
- >700 house environments
- 28 question types (vs. 4 in EQAv1, 6 in MT-EQA)
- 70 possible answers

	Houses	Samples
Train	622	84990
Validation	65	8755
Test	56	7587







Sample videos + questions

'Where is the green rug next to the sofa?'

'Are the computer and the bed the same color?'

'What is the thing next to the tv stand located in the living room?'

New challenges

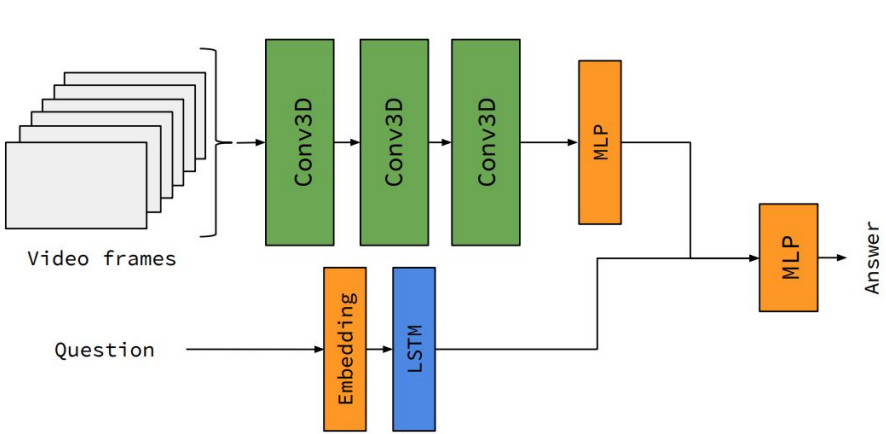
- Rich visual temporal stream of concepts
- NL-like question requiring more varied reasoning capabilities than before
  - Single/multiple object/room existence
  - Object/room counting
  - Object color recognition
  - Object localization
  - Spatial reasoning
  - Object/room size comparison
  - Equality of object attributes

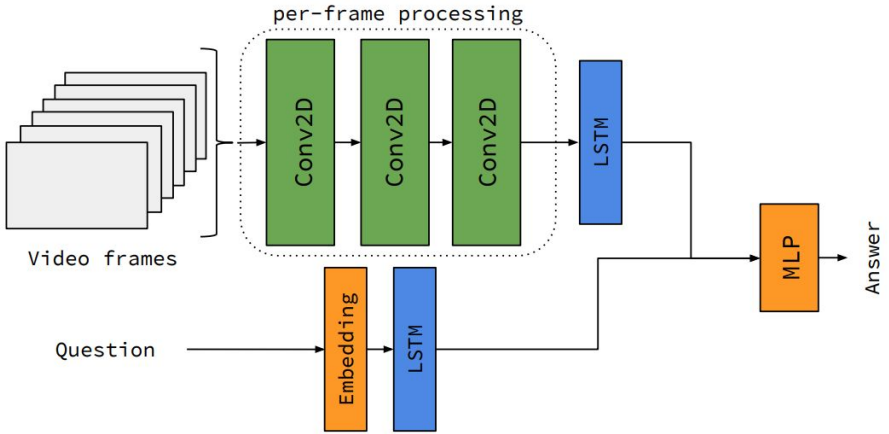
EQAv1 (Q types: 4)	What room is the <OBJ> located in?
	What color is the <OBJ> in the <ROOM>?
VideoNavQA (Q types: 28)	Are both <attr1> <OBJ1> and <attr2> <OBJ2> <color>?
	How many <attr> <OBJ> are in the <ROOM>?
	Is there <art> <attr> <OBJ>?

Approach

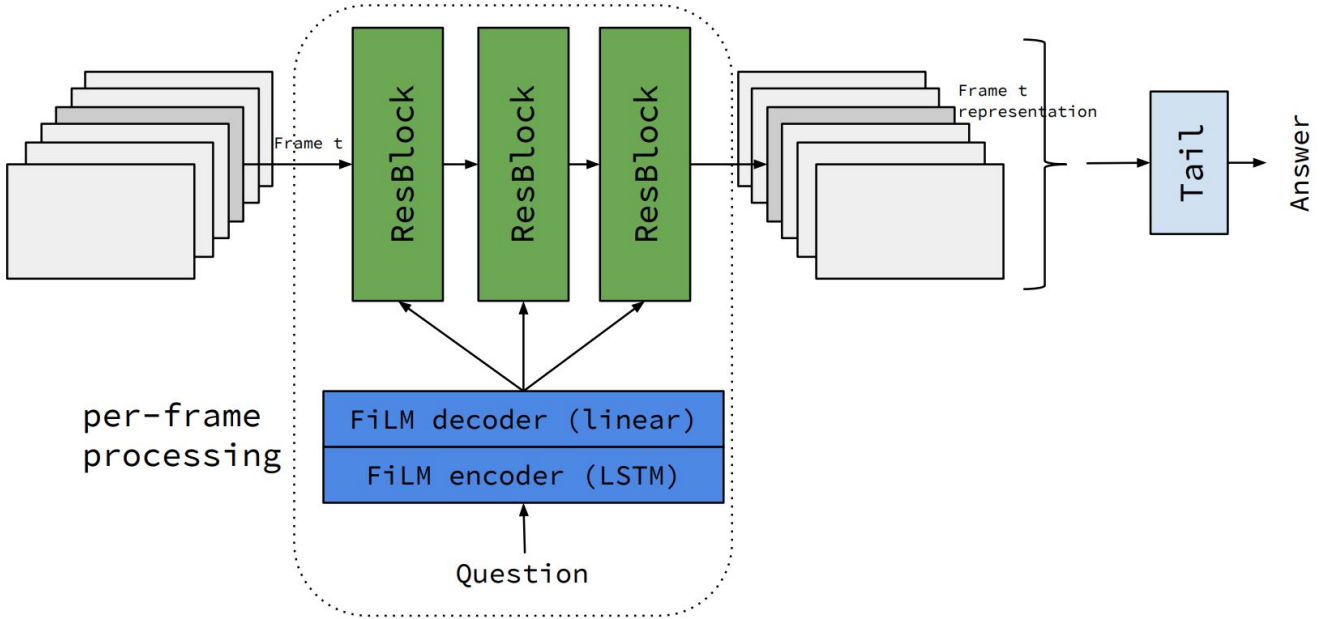
- Leverage VQA paradigms
  - FiLM - language signal used to reweight visual feature maps
  - MAC - sequential reasoning over KB (visual) using language as instruction
- Added difficulty
  - **Temporal** dimension
    - Salient feature selection required during temporal aggregation
  - **Abundant** visual data
    - Object isolation during frame processing

Language+vision models

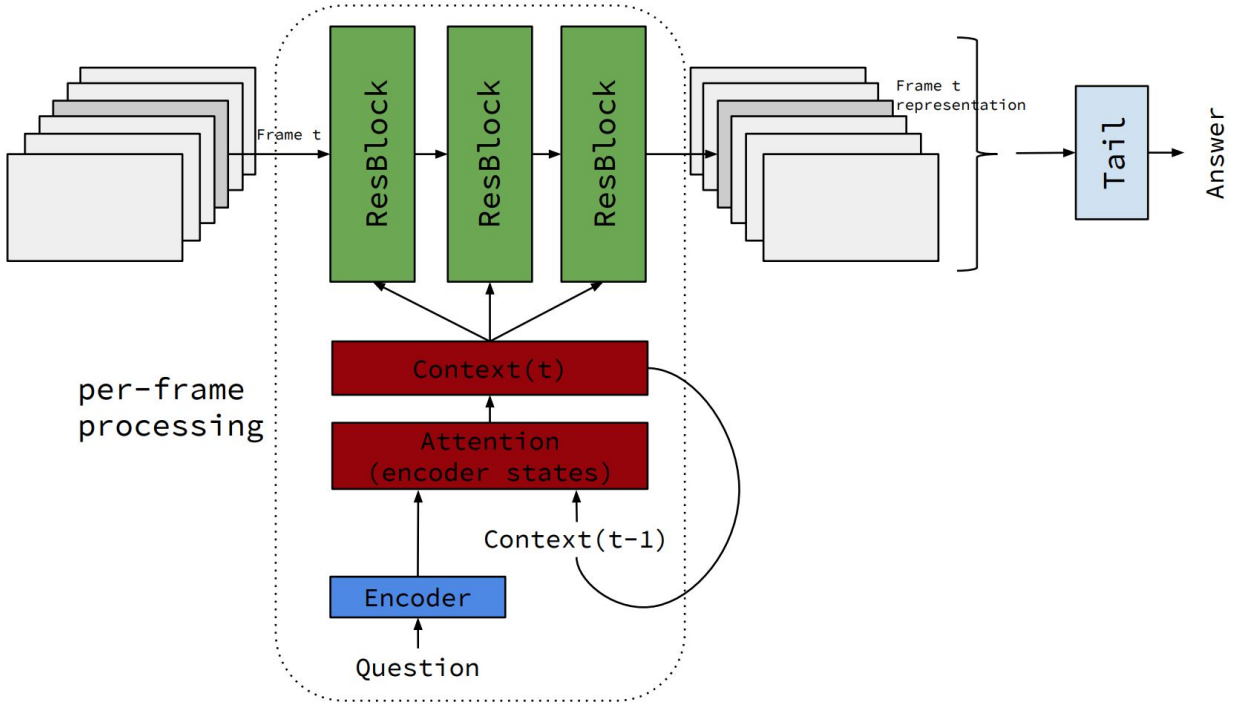




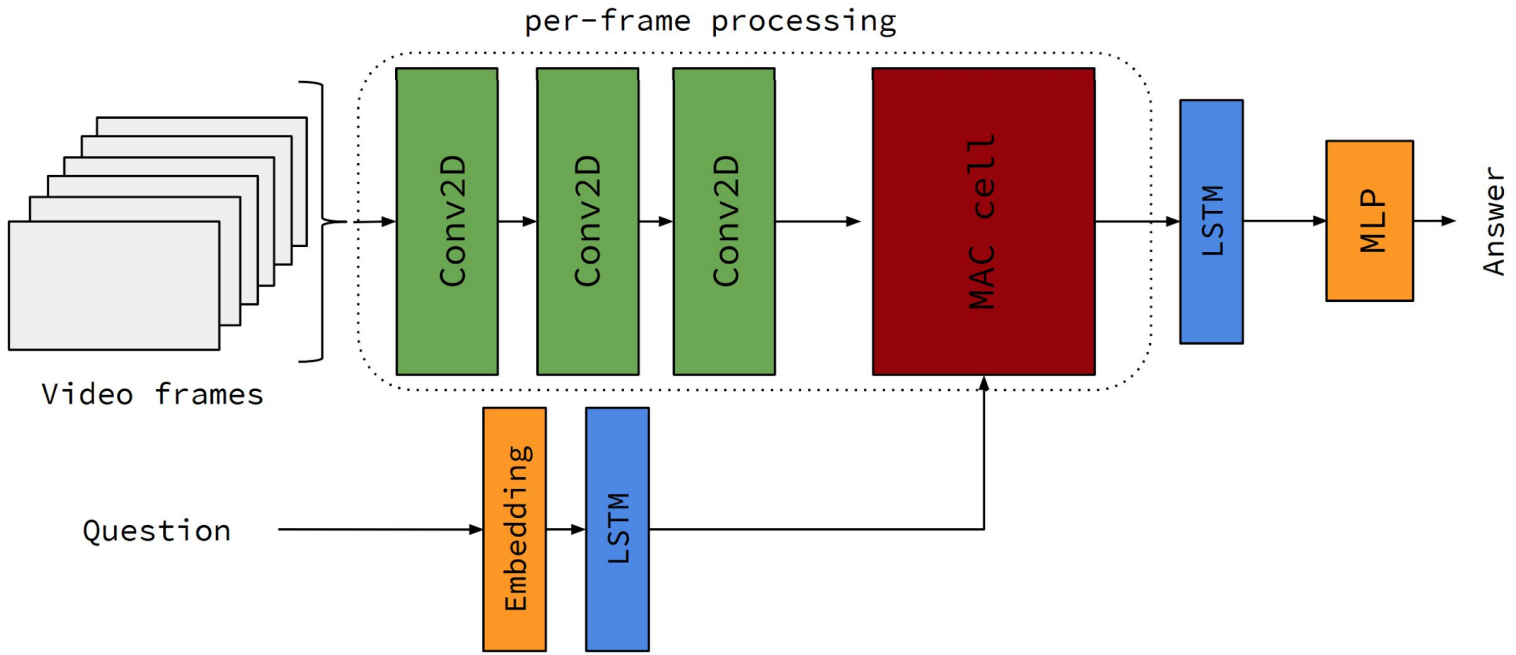
Temporal FiLM



Temporal Multi-hop FiLM



Temporal MAC



Remember blindfold baselines?

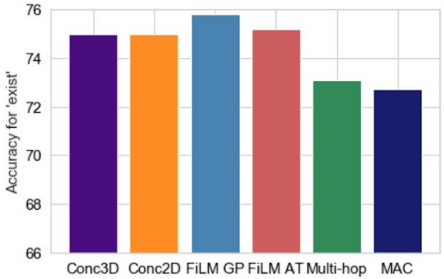
- Biases in the data distribution led to a better-performing “blind” agent!
- Include language-only baselines
  - LSTM
  - BoW
- Place lower bound on VideoNavQA task performance

Initial results

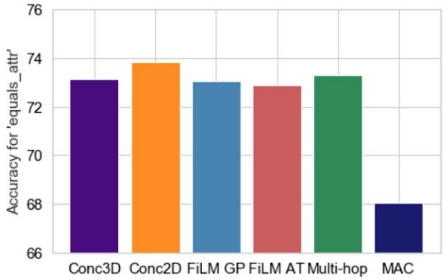
Model	Accuracy All	Yes/No	Other	Num
BoW	49.02	57.67	30.57	40.21
LSTM	56.49	68.36	35.27	38.90
Concat-CNN3D	64.00	72.99	49.12	49.10
Concat-CNN2D	64.47	73.50	49.20	49.59
FiLM-GP	63.79	72.91	47.71	50.00
FiLM-AT	64.08	72.93	49.54	49.26
Temporal multi-hop	63.53	71.81	49.54	50.16
MAC	62.32	69.02	51.37	50.99

Question categories

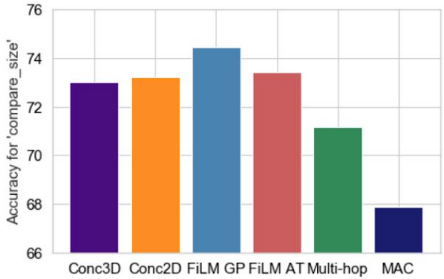
Existence



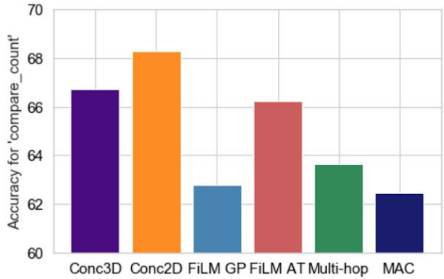
Equality of attributes

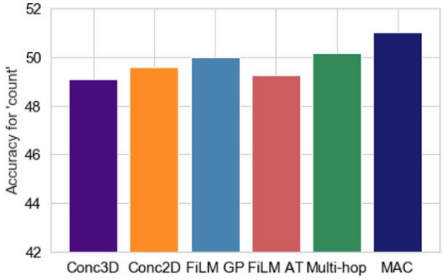


Size comparison

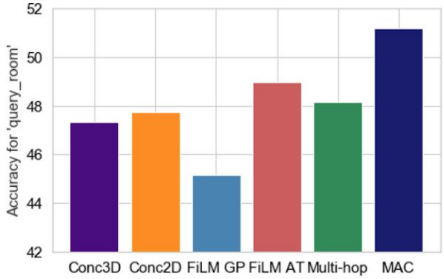


Count comparison

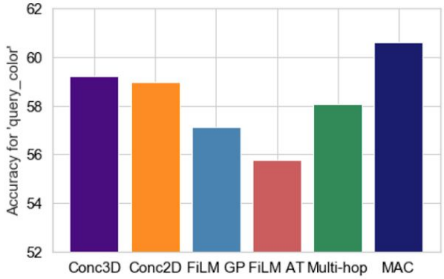




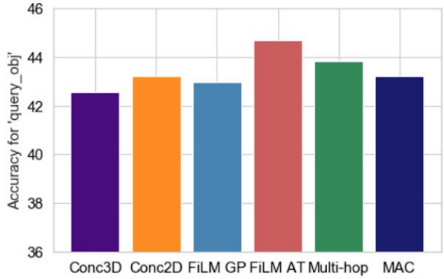
Count



Query room



Query color



Query object

What's next?

- How much can we improve using VQA-style approaches?
- Spatio-temporal graphs?
- Ceiling performance
  - **VideoNavQA provides navigation in the most ideal setting**
  - **Should we aim to reach the same performance on EQA?**
  - **Gradual & measured improvement when integrating navigation**
- Going beyond the "like" in NL-like questions
- Robustness, generalization across environments... :-)

Special thanks

Aaron Courville
Mila

Eugene Belilovsky
Mila

Pietro Liò
University of Cambridge

Ankesh Anand
Mila

Funding: DREAM CDT (PhD), Mila (summer ‘18 internship)
Resources: Mila cluster, University of Cambridge HPC
Dataset hosting: Mila

Thank you!

Paper: arxiv.org/abs/1908.04950

Code and dataset: github.com/catalina17/VideoNavQA

www.cst.cam.ac.uk/~ccc53

@catalinacangea